MLX
Machine Learning Infrastructure

Michael Turmon
Jet Propulsion Laboratory
Exploration Systems Autonomy Section

Ben Bornstein, Robert Granat, Joe Roden, Lucas Scharenbroich

Intelligent Systems Workshop
6 February 2004, Dana Point, California

Summary

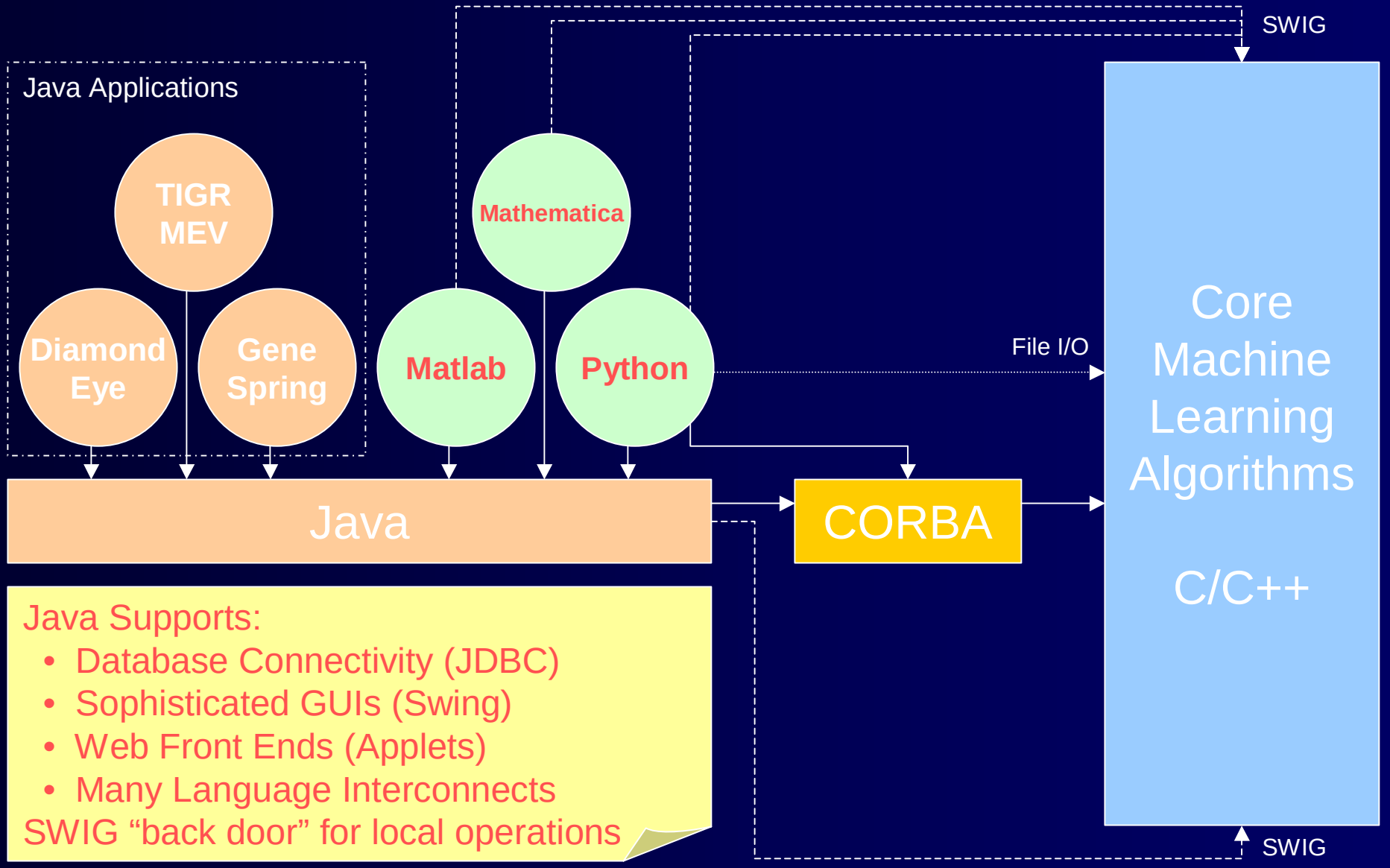
- MLX System Overview
  - Capabilities
  - Architecture
  - Example of use: Clustering
- FY2003 Activity: Spatiotemporal Analysis
  - Object Tracking
  - Track Analysis
- Coming Developments
  - Integration with OODT for DAAC access to AIRS data
  - Integration in SERVOgrid for Solid Earth time series

MLX Capabilities

MLX: A software infrastructure for discovery of knowledge in terabyte-scale data sets.

- Allow investigators to understand large data sets with minimal up-front development costs
  - Factor out common aspects of project or mission-scale data analysis; we are developing a DAAC interface for *data access*
  - Provide a portfolio of basic algorithms helpful in any machine learning effort: *methods database*
  - Harness plentiful computational resources with frameworks for large-scale tuning, parameter sweeps, cross-validation
- Provide new capabilities to analysts of massive datasets
  - Time series analysis: decomposition, segmentation, classification, novelty and outlier detection
  - Image classification, object recognition and tracking

MLX Architecture



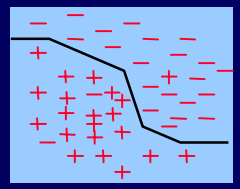
Java Supports:

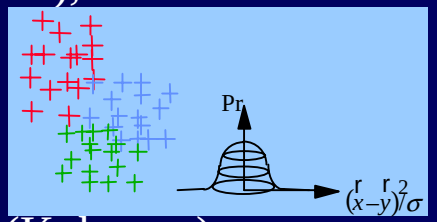
- Database Connectivity (JDBC)
- Sophisticated GUIs (Swing)
- Web Front Ends (Applets)
- Many Language Interconnects

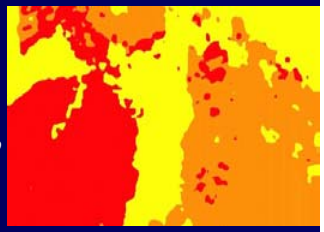
SWIG “back door” for local operations

Machine Learning Tasks

- Multiple learning methods *(methods database)*
  - Supervised
    - Discriminant analysis, ANN (artificial neural networks), SVM (support vector machines)
  - Unsupervised classification
    - Hierarchical decomposition: TSplit, SClust, XClust
    - Optimization: k-means, EM/normal mixtures, SOM (Kohonen)
  - Spatiotemporal algorithms
    - MRF (Markov Random Field), HMM (Hidden Markov Models), Kalman filters



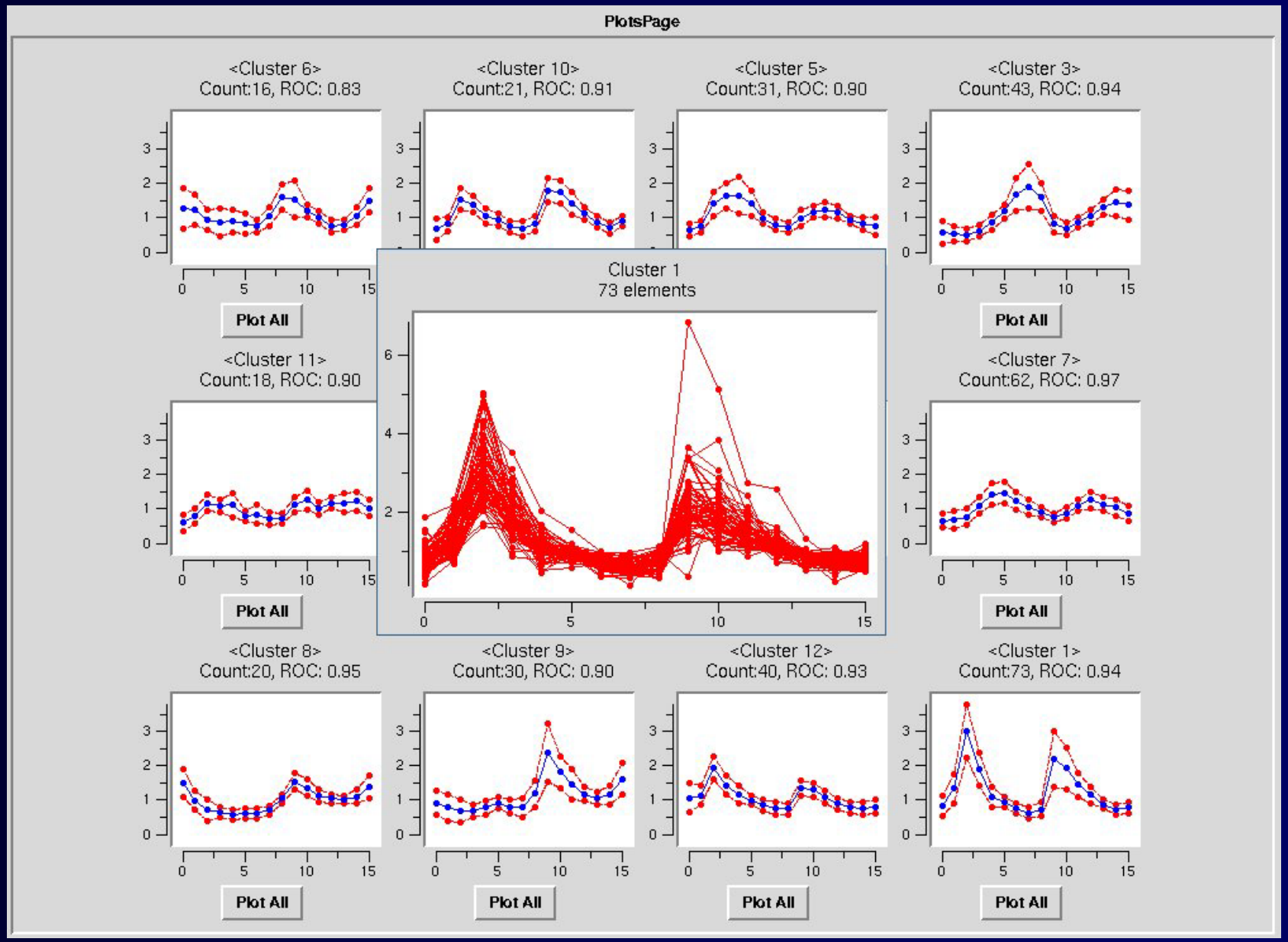




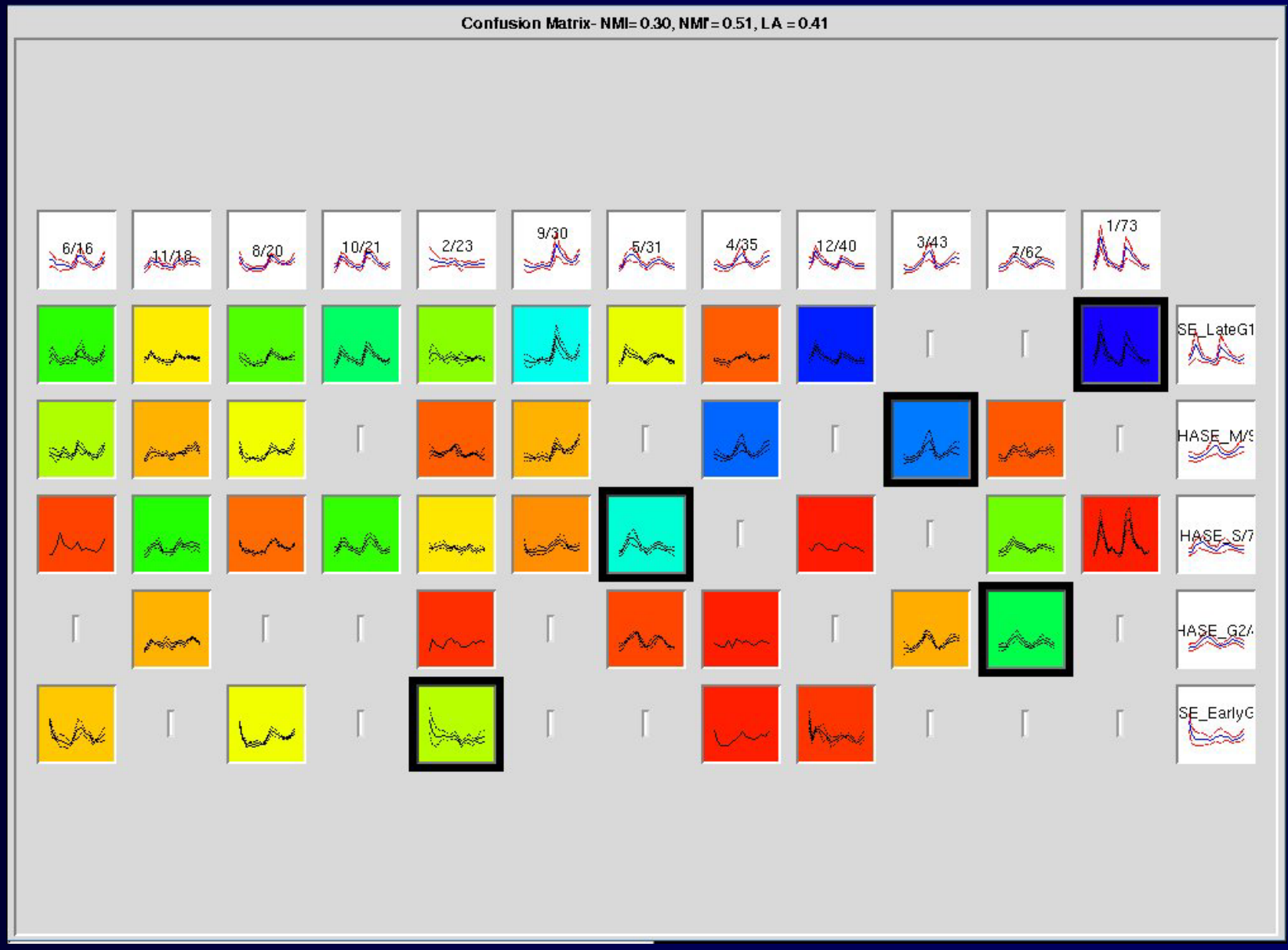
- Comparison and combination
  - Cross-validation, Monte Carlo CV, bootstrapping
  - ROC (Receiver Operating Characteristic), Linear Assignment (cluster matching), Normalized Mutual Information
- Multiple application enivronments
  - C/C++, Java, Python, Matlab, Mathematica, notebooks, etc.

Example 1/3: Visualize Clusterings

California Institute of Technology
M. Turmon / MLX

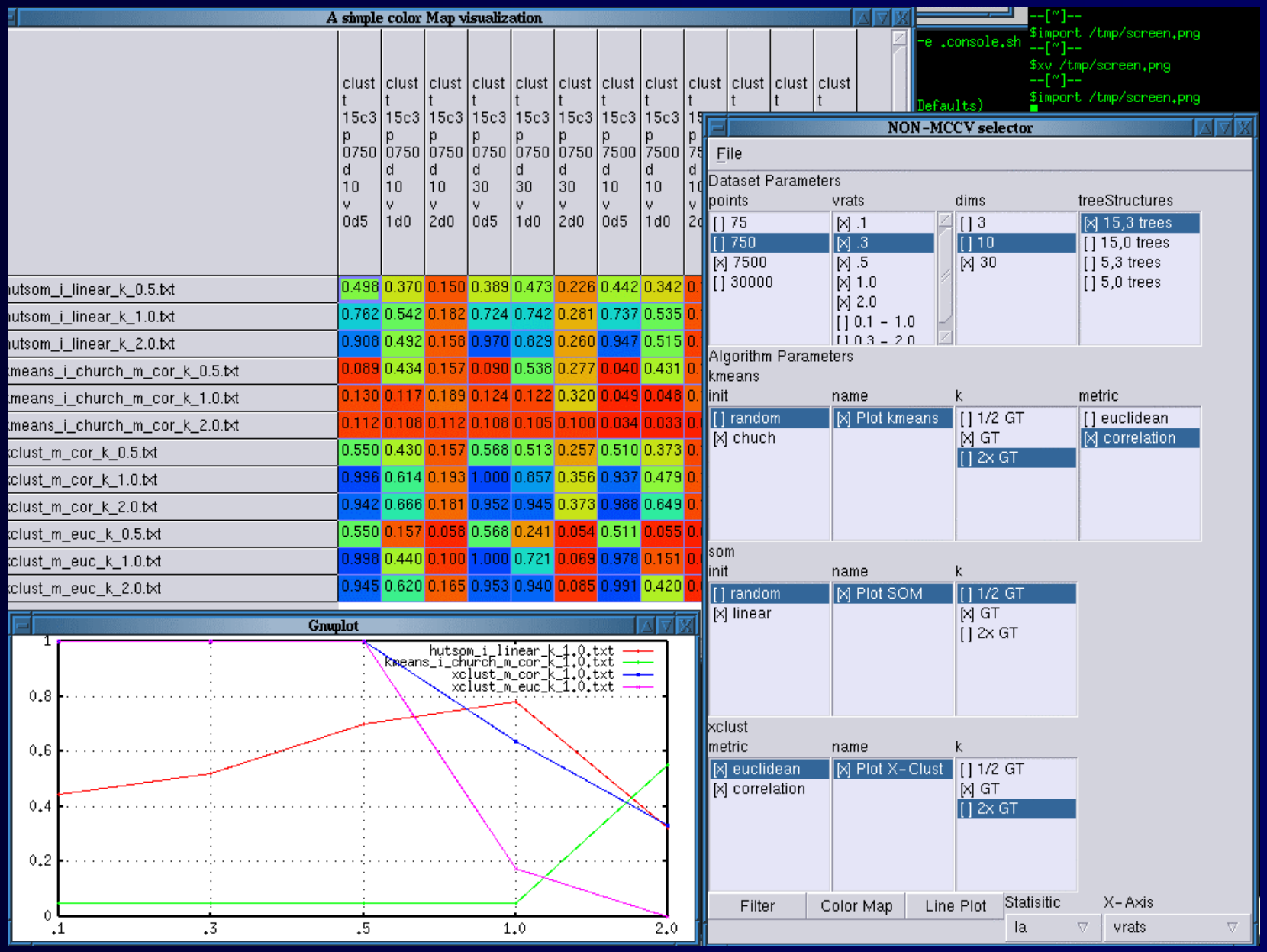


Example 2/3: Compare Clusterings



Example 3/3: Compare Algorithms

California Institute of Technology
M. Turmon / MLX



Object Recognition and Tracking

M. Turmon / MLX

Goal: Understand great volumes of spatiotemporal data in directly informative terms; move from pixels to objects.

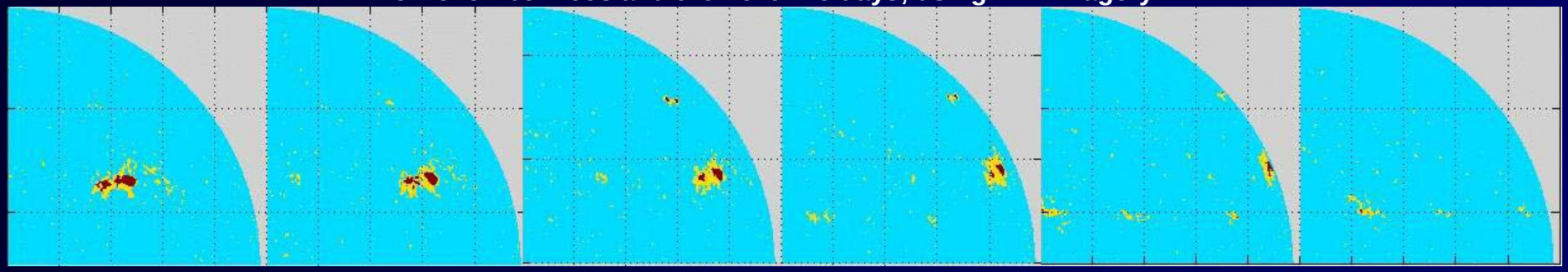
Identification → Tracking → Trajectory analysis

Identification: Find the objects in multispectral science images

Tracking: Link identified objects in series of images

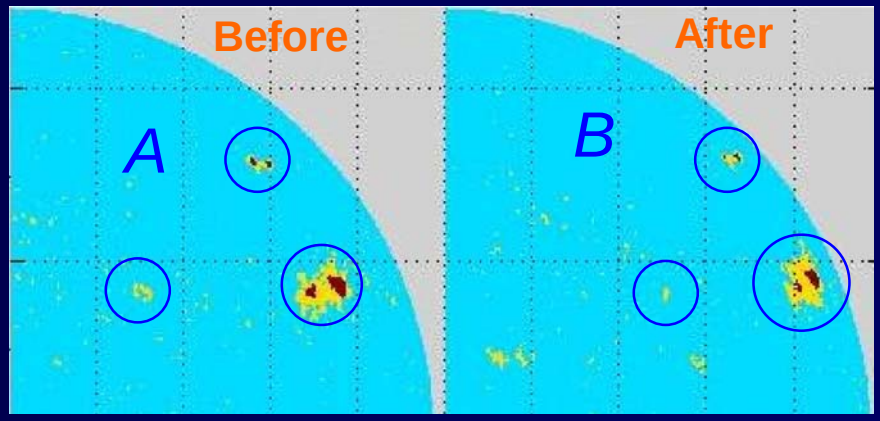
Trajectory Analysis: Model and classify object tracks

Sunspot and Facula Regions in a Solar Quadrant
15 November 1998 and the next five days; using MDI imagery



Object Tracking Methods

- Generic solution is feasible for many problems
- Associate objects in before and after images
- Correlation-based tracker
  - Motion model: deterministic drift plus stochastic uncertainty
  - Sunspots, clouds, cyclones: have motion and correlation on the sphere
  - Correlation measure between *a* in *A* and *b* in *B* is *D(a,b)*
- Solve assignment problem to match *A* up to *A'*:



$$\max_{P} \sum_{a \in A} \sum_{b \in B} P(a,b)\, D(a,b)$$

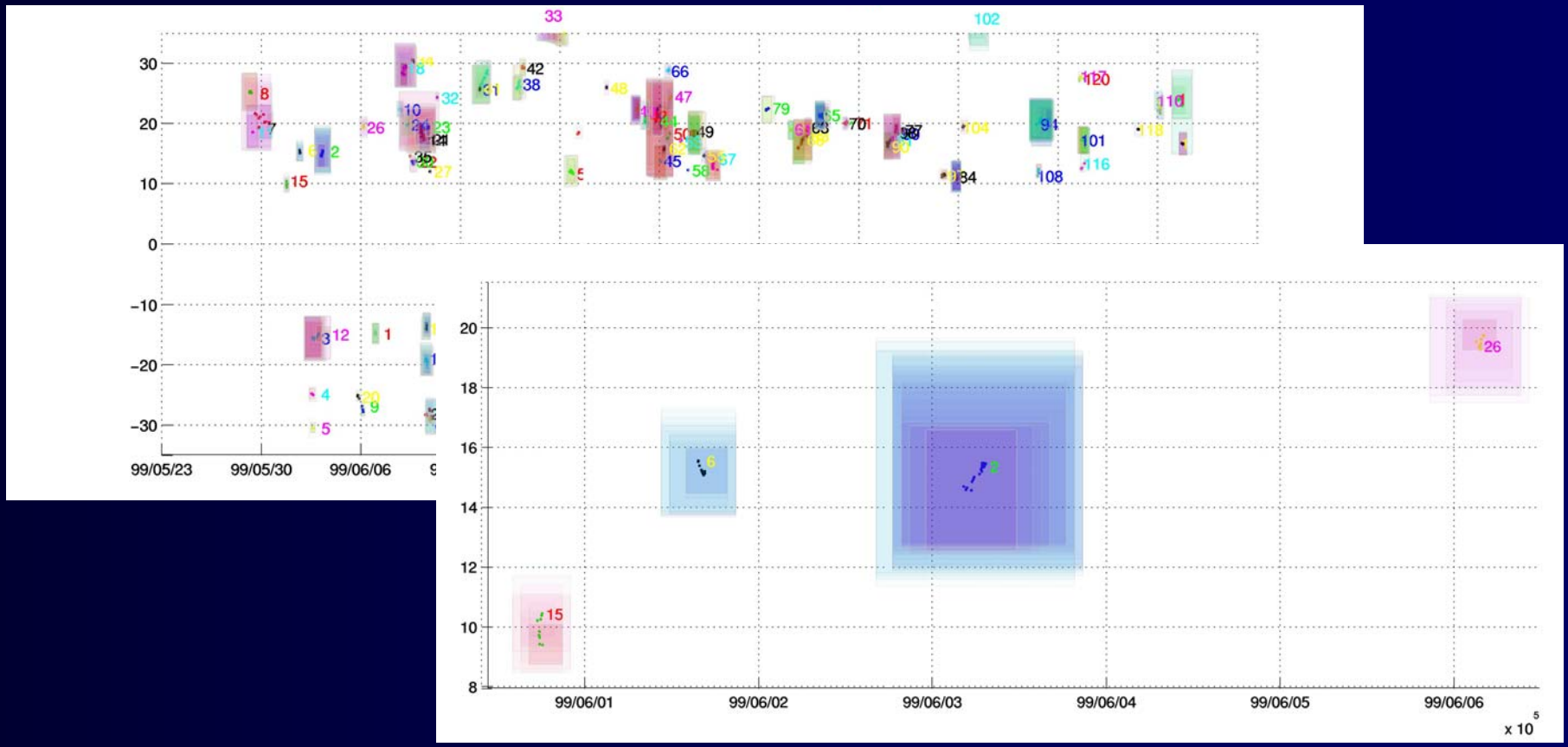
with *P* a permutation matrix

Solution by linear programming

- For our applications, key is to get deterministic drift correct

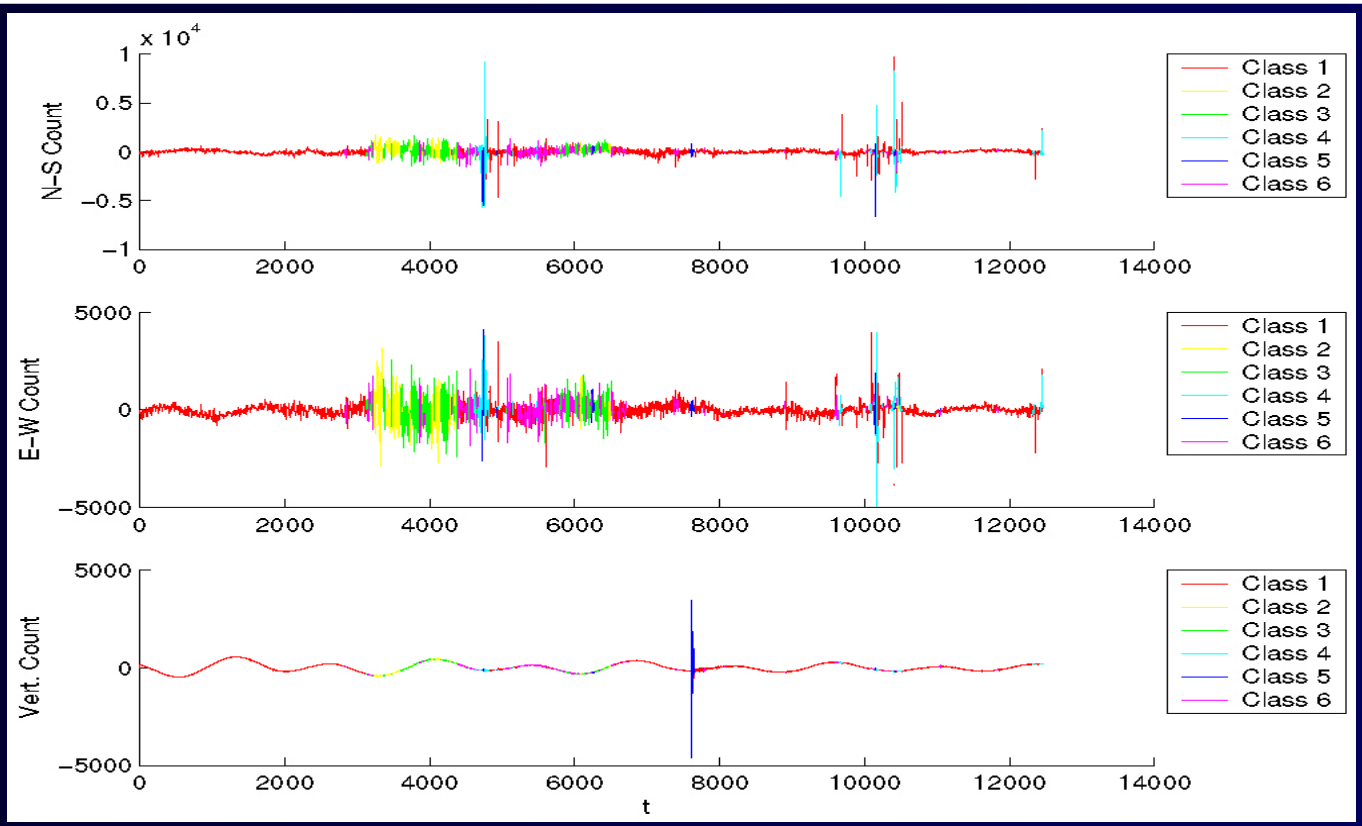
Time Series: Trajectory Results

M. Turmon / MLX



- Coordinates of tracked sunspots (June-July 1999 shown above)
- Model, understand, and classify these trajectories
- Other examples: GPS sensors, cyclone evolution history

Path Analysis: Segmentation



Seismograph (3D), about two days

Preprocessed by Fourier windowing (spectrogram)

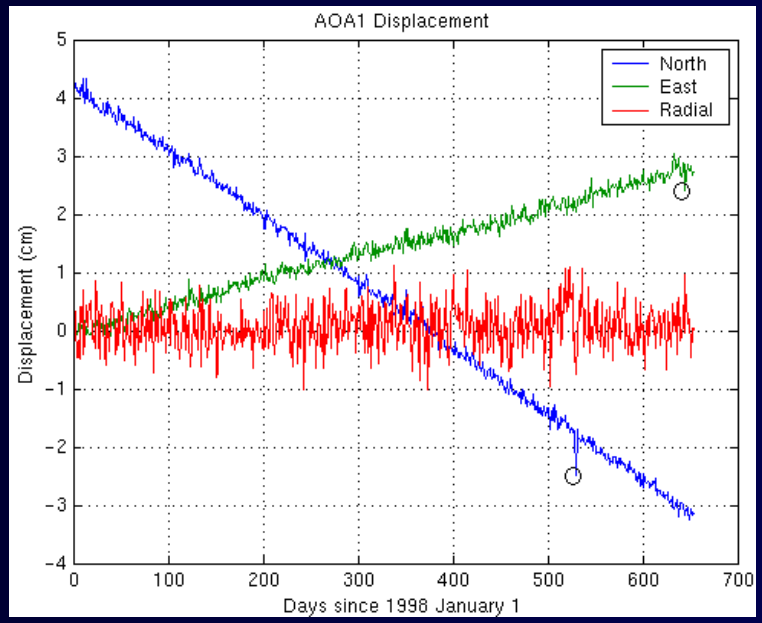
Then divided by HMM into 6 classes

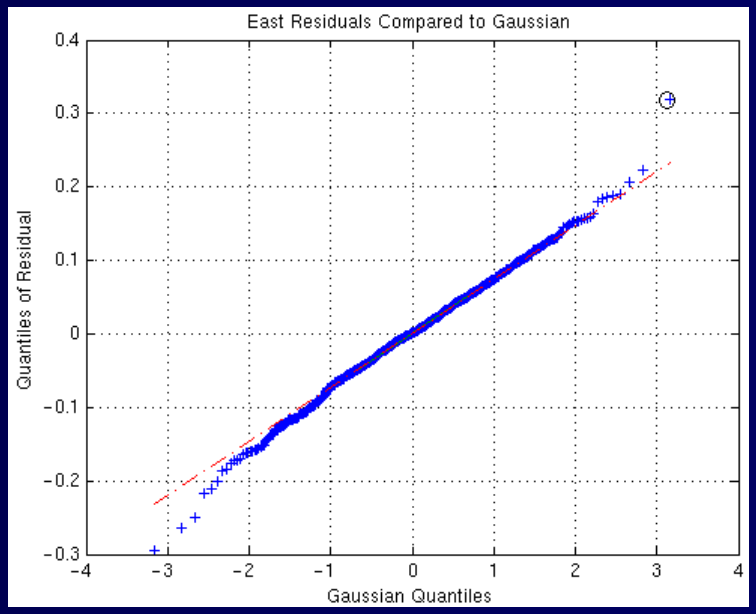
Unexplained “fuzz” separated out near t = 4000; transient identified at t = 7500

- Very long signals with vastly different “modes” are difficult for human analysts to visualize, leading to gaps in analysis
- HMM segmentation provides an automatic way to focus attention on the most interesting parts of the trajectory

Path Analysis: Novelty Detection

M. Turmon / MLX





- Time series of GPS sensor locations to ~mm precision (in 3D)
- Fit by two-state Kalman filter (generalization of AR-2)
  - Residuals highlight two events above noise (1 shown on RHS)
- Trainable signal modeling engine for many learning tasks
  - Train signal models automatically based on observed data
    - Less sensitive to slow fluctuations than Fourier methods
  - Novelty or outlier detection: find signals not present at original training

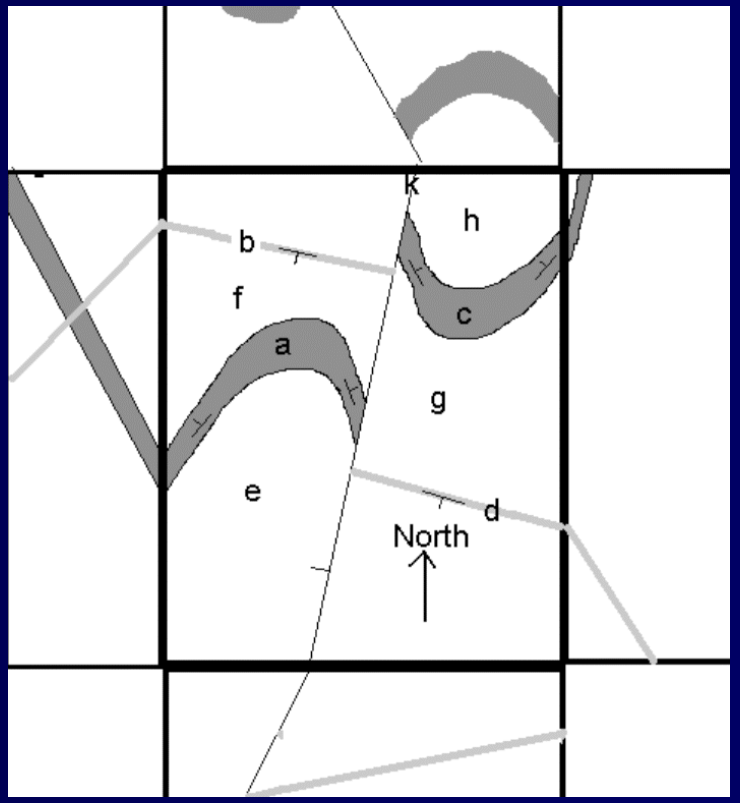
Path Analysis: Geological Constraints

- Real geological structures have movement constraints among lithological units
  - E.g., E, F, G at right do not move independently
- If known, this translates into model constraints
  - (…and illustrates the value of models…)
  - In particular, the Kalman filter

$$P(u(t+1) \mid u(t)) = N(Cu(t), \Sigma)$$
$$P(z(t) \mid u(t)) = N(Bu(t), \Sigma_z)$$

should inherit constraints on its motion feedback matrix C

- When we have this information, we can use it to improve model fits
  - If unknown, related motions may be discovered through path analysis



MLX Planned Activities

- Extend MLX Core
  - Cluster/classify using partially labeled data
  - Robust parallelism and scheduling support (OpenPBS etc.)
- Create DAAC interface (Crichton/OODT; Fetzer et al./AIRS)
  - Demonstrate remote data access by MLX data mining components on the Goddard DAAC using OODT accessors
  - Detect and track moisture flares in AIRS data
- Geophysics integration (A. Donnellan, K. Hurst/SERVOGrid)
  - Classify GPS deformation signals
  - Combine GPS and broadband seismic data
  - Use as tool to link crustal motion data and fault models

Seeking other collaborations with domain scientists